Summary of

Standards for Educational and Psychological Testing

Courtesy of the Test Validation & Construction Unit

Introduction

This summary of the *Standards for Educational and Psychological Testing* is intended to provide a brief overview of the standards essential to the field of testing. The intent of the *Standards* is to promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices. This summary should be used in conjunction with the full text of the *Standards for Educational and Psychological Testing* to address specific selection-related queries.

The *Standards* were prepared by the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

The *Standards* are not in and of themselves legislation or law; however, they should be considered and implemented, when appropriate, by individuals in the field of testing. It is vital to utilize these standards, since the improper use of tests can cause considerable harm to test takers and other parties affected by test-based decisions. These doctrines provide assessment professionals with guidelines for the evaluation, development, and use of testing instruments.

History and Purpose of the *Standards for Educational and Psychological Testing*

The purpose of publishing the *Standards* is to provide criteria for the evaluation of tests, testing practices, and the effects of test use. Although the evaluation of the appropriateness of a test or testing application should depend heavily on professional judgement, the *Standards* provide a frame of reference to assure that relevant issues are addressed.

It is hoped that all professional test developers, sponsors, publishers, and users will adopt the *Standards* and encourage others to do so. Overall, the *Standards* advocate that, within feasible limits, the relevant technical information be made available so that those involved in policy debate may be fully informed.

Organization of the *Standards for Educational and Psychological Testing*

The *Standards* are organized into three major sections. The document begins with a series of chapters devoted to the test development process, which focus primarily on the responsibilities of test developers. The *Standards* continue with chapters that address specific uses and applications of the standards, which focus primarily on the responsibilities of test users. Each chapter within the major sections begins with an introductory text, providing an explanatory background for the standards that follow.

Part I of the *Standards*: *Test Construction, Evaluation, and Documentation.*

1. Standards for Validity

The *Standards* refer to validity as the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is the most fundamental consideration in developing and evaluating tests. Professional judgement guides decisions regarding the specific forms of evidence that can best support the intended interpretation and use.

> **Standard 1.4**
> **If a test is used in a way that has not been validated, it is incumbent on the user to justify the new use, collecting new evidence if necessary.** (see pg. 18)

The *Standards* outline the following sources of validity evidence: evidence based on test content, response processes, internal structure, relations to other variables such as convergent and discriminant evidence, test criterion relationships, validity generalization, and finally evidence based on consequences of testing.

2. Reliability and Errors of Measurement

Reliability refers to the consistency of measurements when the testing procedure is repeated on a population of individuals or groups. However, no single examinee is completely consistent, and in some instances, because of subjectivity in the scoring process, an individual's obtained score and the average score of a group will always reflect at least a small amount of measurement error. Information about measurement error is essential to the proper evaluation and use of a test instrument.

The standard error of measurement is typically more relevant once a measurement procedure has been adopted and interpretation of scores has become a user's primary concern. No test developer is exempt from the responsibility of investigating test reliability as fully as practical considerations permit.

> **Standard 2.1**
> **For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors and standard errors of measurement or test information functions should be reported.** (see pg. 31)

3. Test Development and Revision

Test development is the process of producing a measure of some aspect of an individual's knowledge, skill, ability, interests, attitudes, or other characteristics by developing items and combining them to form a test, according to a specified plan. Test development also

includes specifying conditions for administering the test, determining procedures for scoring the test performance, and reporting the scores to test users.

The chapter in the *Standards* which addresses test development and revision, focuses on stating the purpose of the test, defining a framework for the test, developing test specifications, developing and evaluating items and their associated scoring procedures, assembling the test, and revising the test.

The following two standards are important to keep in mind during test development and revision.

> **Standard 3.8**
> **When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s) should be as representative as possible of the population(s) for which the test is intended.** (see pg. 44)

> **Standard 3.13**
> **When a test score is derived from the differential weighting of items, the test developer should document the rationale and process used to develop, review, and assign item weights. When item weights are obtained based on empirical data, the sample used for obtaining item weights should be sufficiently large and representative of the population for which the test is intended.** (see pg. 46)

4. Scaling, Norming, and Score Comparison

Scale scores may aid in the interpretation of scores by indicating how a given score compares to those of other test takers, by enhancing the comparability of scores obtained using different forms of a test or in other ways. The validity of norm-referenced interpretations depends in part on the appropriateness of the reference group to which tests scores are compared. It is important that norms be based on a technically sound, representative, scientific sample of sufficient size.

In reference to score comparison, the *Standards* clearly state:

> **Standard 4.15**
> **When additional test forms are created by taking a subset of the items in an existing test form or by rearranging its items and there is sound reason to believe that scores on these forms may be influenced by item context effects, evidence should be provided that there is no undue distortion of norms for the different versions or of score linkages between them.** (see pg. 58)

Cut Scores
A critical step in the development and use of some tests is to establish one or more cut points dividing the score range to partition the distribution of scores into categories. An employer may determine a cut score to screen potential employees or promote current employees.

Where the results of the standard-setting process have highly significant consequences, and especially where large numbers of examinees are involved, those responsible for establishing cut scores should be concerned that the process by which cut scores are determined be clearly documented and defensible (pg. 53)

> **Standard 4.19**
> **When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented.** (see pg. 59)
>
> **Standard 4.20**
> **When feasible, cut scores defining categories with distinct substantive interpretations should be established on the basis of sound empirical data concerning the relation of test performance to relevant criteria.**
> (see pg. 60)

5. Test Administration, Scoring, and Reporting

The *Standards* explain that the usefulness and interpretability of test scores require that the directions to examinees, testing conditions, and scoring procedures follow the same detailed procedures. When these steps are taken, the test is said to be standardized, without such, the accuracy and comparability of score interpretations would be reduced.

6. Supporting Documentation for Tests

The *Standards* describe that test documents need to include enough information to allow test users and reviewers to determine the appropriateness of the test for its intended purposes.

A test's documentation typically specifies the nature of the test; its intended use; the process involved in the test's development; technical information related to scoring, interpretation, and evidence of validity and reliability; scaling and norming if appropriate to the instrument; and guidelines for test administration and interpretation.

Part II of the Standards: *Fairness in Testing*

1. Standards on Fairness and Bias

The *Standards* focus on the aspects of fairness and testing that are customarily the responsibility of those who make, use, and interpret tests, which are characterized by

some level of professional; and technical consensus. It does not examine the very broad issues related to regulations, statutes, and case law that govern test use and the remedies for harmful practice.

The *Standards* describe fairness in the following four principle ways in which the term fairness should be used: Fairness as a Lack of Bias; Fairness as Equitable Treatment in the Testing Process; Fairness as Equality in Outcomes of Testing; and Fairness as Opportunity to Learn.

The *Standards* describe the term bias as construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees. Likewise, two main sources of bias are identified: Content-Related Sources of Bias and Response-Related Sources of Bias.

2. The Rights and Responsibilities of Test Takers

Fairness issues unique to the interests of the individual test taker, which reflects widely accepted principles in the field of measurement are also addressed. The responsibilities of test takers concerning test security, their access to test results, and their rights when irregularities in their testing are claimed are discussed.

3. Testing Individuals of Diverse Linguistic Backgrounds

Any test that employs language is, in part, measuring a test taker's language skills. It is important to consider language background in developing, selecting, administering, and interpreting test performance. In addition, in cases where a language-oriented test is inappropriate due to the test taker's limited proficiency in that language, a non-verbal test may be a suitable alternative.

This section of the *Standards* explores the following topics: Test Translation, Adaptation, and Modification; Issues of Equivalence; Language Proficiency Testing; Testing Bilingual Individuals; Administration and Examiner Variables; Use of Interpreters in Testing; and Cultural Differences and Individual Testing.

4. Testing Individuals with Disabilities

Although the *Standards* focus on technical and professional issues regarding the testing of individuals with disabilities, test developers and users are also encouraged to become familiar with federal, state, and local laws, and court and administrative rulings that regulate the testing and assessment of individuals with disabilities.

However, the *Standards* do address issues regarding appropriate accommodations when testing individuals with disabilities, strategies of test modification, using modifications in different testing contexts, and reporting scores on modified tests.

Part III of the standards: *Testing Applications*

1. Standards Involving General Responsibilities of Test Users

Test users are referred to as the group of professionals who actively participate in the interpretation and use of test results. In selecting a test and interpreting a test score, the test user is expected to have a clear understanding of the purposes of the testing and its probable consequences.

> **Standard 11.2**
> **When a test is to be used for a purpose for which little or no documentation is available, the user is responsible for obtaining evidence of the test's validity and reliability for this purpose.** (see pg. 113)
>
> **Standard 11.19**
> **When a test user contemplates an approved change in test format, mode of administration, instructions, or the language used in administering the test, the user should have a sound rationale for concluding that validity, reliability, and appropriateness of norms will not be compromised.** (see pg. 117)

Testing decisions can be justified by documenting that test uses and score interpretations are supported by measurement authorities for the given purpose. In addition, the inferences drawn should be validated for use for the given population, and the results should be used in conjunction with other information.

2. Psychological Testing and Assessment

This chapter of the *Standards* addresses issues important to professionals who use psychological testing with their clients. The topics covered include test selection and administration, test interpretation, collateral information used in psychological testing, types of tests, and purposes of testing.

The types of psychological tests described include: Cognitive and Neuropsychological Testing; Social, Adaptive, and Problem Behavior Testing; Family and Couples Testing; Personality Testing; Vocational Testing; Interest Inventories; Work Values Inventories; and Measures of Career Development, Maturity, and Indecision. The purposes of psychological testing are outlined as Testing for Diagnosis; Testing for Intervention Planning and Outcome Evaluation; Testing for Judicial and Governmental decisions; and Testing for Personal Awareness, Growth, and Action.

3. Educational Testing and Assessment

Within this context, the *Standards* are concerned with testing in formal educational settings from kindergarten through post-graduate training. In this section, three broad areas of educational testing are considered, which encompass various given purposes of educational testing. They are: (a) routine school, district, state, or other system-wide

testing programs, (b) testing for selection in higher education, and (c) individualized and special needs testing.

4. Testing in Employment and Credentialing

The *Standards* describe employment testing as being carried out by organizations for purposes of employee selection, promotion, or placement. These three purposes all focus on the prediction of future job behaviors, with the goal of influencing organizational outcomes, such as efficiency, growth, productivity, and employee motivation and satisfaction. The *Standards* address testing for licensure and certification, with a focus on the applicant's current skill or competency in a specified domain.

> **Standard 14.1**
> **Prior to development and Implementation of an employment test, a clear statement of the objective of testing should be made. The subsequent validation effort should be designed to determine how well the objective has been achieved.** (see pg. 158)

The *Standards* address the following contextual features: (a) internal versus external candidate pool, (b) untrained versus specialized jobs, (c) short-term vs. long-term focus, (d) screen in versus screen out, (e) mechanical versus judgmental decision, (f) ongoing versus one-time use of a test, (g) fixed applicant pool versus continuous flow, (h) small versus large sample size, and (i) size of applicant pool, relative to number of job openings.

Employment testing is highly influenced by the content of test use, which covers various domains of knowledge, skills, abilities, traits, dispositions, and values. The Context is also very influential to employment testing. The following two standards address test content:

> **Standard 14.8**
> **Evidence of validity based on test content requires a thorough and explicit definition of the content domain of interest. For selection, classification, and promotion, the characterization of the domain should be based on job analysis.** (see pg. 160)

> **Standard 14.9**
> **When evidence of validity based on test content is a primary source of validity evidence in support of the use of a test in selection or promotion, a close link between test content and job content should be demonstrated.** (see pg. 160)

The fundamental inference typically drawn from test scores in an employment setting is a prediction, whereby the employer wants to make an inference from test results to future

job performance. The validation process in employment settings involves the gathering and evaluation of evidence relevant to sustaining or challenging this inference.

The empirical link between the predictor measure (the test) and the criterion measure (job behavior or outcome of interest) must be supplemented by evidence of the relevance of the criterion measure to the criterion construct domain to complete the linkage between the test and the criterion construct domain.

> **Standard 14.4**
> **When empirical evidence of predictor-criterion relationships is part of the pattern of evidence used to support test use, the criterion measure(s) used should reflect the criterion construct domain of interest to the organization. All criteria used should represent important work behaviors or work outputs, on the job or in job relevant training, as indicated by an appropriate review of information about the job.** (see pg. 159)

> **Standard 14.5**
> **… empirical studies of predictor-criterion relationships should identify contaminants and artifacts that may have influenced study findings, such as error of measurement, range restriction, and the effects of missing data.** (see pg. 159)

Regarding testing in professional and occupational credentialing, the *Standards* state that licensing requirements are imposed by state and local governments to ensure that those licensed possess knowledge and skills in sufficient degree to perform important occupational activities safely and effectively. Tests used in credentialing should be designed to determine whether the essential knowledge and skills of a specified domain have been mastered by the candidate.

> **Standard 14.17**
> **The level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test.** (see pg. 162)

Validation of credentialing tests depends mainly on content-related evidence, mainly in the form of judgements that the test adequately represents the content domain of the occupation or specialty being considered.

5. Testing in Program Evaluation and Public Policy

Tests are also widely used in program evaluation, which assesses the needs, implementation, effectiveness, and value of a program, and in public policy decision making. Test results can be utilized as a source of evidence for the initiation, continuation, modification, termination, or expansion of various programs and policies.

The *Standards* recognize that it is important to evaluate any proposed test in terms of its relevance to the goals of the program or policy and/or to the particular question its use will address.

> **Standard 15.1**
> **When the same test is designed or used to serve multiple purposes, evidence of technical quality for each purpose should be provided.**
> (see pg. 167)